

HyperFusion: Hierarchical Multimodal Ensemble Learning for Social Media Popularity Prediction

Liliang Ye
Huazhong University of Science and
Technology
Wuhan, China
yll@hust.edu.cn

Yunyao Zhang
Huazhong University of Science and
Technology
Wuhan, China
ikoyun@hust.edu.cn

Yafeng Wu
Huazhong University of Science and
Technology
Wuhan, China
wyf2024@hust.edu.cn

Yi-Ping Phoebe Chen
La Trobe University
Melbourne, Australia
phoebe.chen@latrobe.edu.au

Junqing Yu
Huazhong University of Science and
Technology
Wuhan, China
yjqing@hust.edu.cn

Wei Yang
Huazhong University of Science and
Technology
Wuhan, China
weiyangcs@hust.edu.cn

Zikai Song*
Huazhong University of Science and
Technology
Wuhan, China
skyesong@hust.edu.cn

Abstract

Social media popularity prediction plays a crucial role in content optimization, marketing strategies, and user engagement enhancement across digital platforms. However, predicting post popularity remains challenging due to the complex interplay between visual, textual, temporal, and user behavioral factors. This paper presents HyperFusion, a hierarchical multimodal ensemble learning framework for social media popularity prediction. Our approach employs a three-tier fusion architecture that progressively integrates features across abstraction levels: visual representations from CLIP encoders, textual embeddings from transformer models, and temporal-spatial metadata with user characteristics. The framework implements a hierarchical ensemble strategy combining CatBoost, TabNet, and custom multi-layer perceptrons. To address limited labeled data, we propose a two-stage training methodology with pseudo-labeling and iterative refinement. We introduce novel cross-modal similarity measures and hierarchical clustering features that capture inter-modal dependencies. Experimental results demonstrate that HyperFusion achieves competitive performance on the SMP challenge dataset. Our team achieved third place in the SMP Challenge 2025 (Image Track). The source code is available at <https://anonymous.4open.science/r/SMPDImage>.

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '25, Dublin, Ireland

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
<https://doi.org/XXXXXXX.XXXXXXX>

CCS Concepts

• Information systems → Multimedia information systems.

Keywords

Social Media Popularity Prediction, Multimodal Machine Learning, Feature Construction

ACM Reference Format:

Liliang Ye, Yunyao Zhang, Yafeng Wu, Yi-Ping Phoebe Chen, Junqing Yu, Wei Yang, and Zikai Song. 2025. HyperFusion: Hierarchical Multimodal Ensemble Learning for Social Media Popularity Prediction. In *Proceedings of ACM International Conference on Multimedia 2025 (MM '25)*. ACM, New York, NY, USA, 8 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

The exponential growth of social media platforms has fundamentally transformed digital communication, with billions of posts shared daily on diverse platforms around the world [24]. Predicting the popularity of social media content has emerged as a critical research area with profound implications for content optimization, marketing strategies, and algorithmic recommendation systems [11]. The complexity of this task stems from the intricate interplay between multiple heterogeneous factors, including visual aesthetics, textual semantics, temporal dynamics, spatial context, and user behavioral patterns. Early research [14, 21, 31, 33, 34, 36] primarily relied on handcrafted features derived from textual content and basic metadata such as posting timestamps and user demographics. While these approaches provided foundational insights, they often failed to capture the rich semantic information embedded in visual content and the complex cross-modal relationships that significantly influence user engagement. The advent of deep learning has revolutionized this field, enabling researchers to leverage sophisticated neural architectures for extracting meaningful representations from multimodal data [3, 4, 17, 38]. Recent advances

have demonstrated the effectiveness of vision-language models, particularly CLIP [23], in bridging the semantic gap between visual and textual modalities. However, existing approaches often treat different modalities independently or employ simple concatenation strategies for feature fusion, potentially overlooking the hierarchical nature of feature interactions and the varying modality importance across contexts.

To address these limitations, this paper introduces HyperFusion, a novel hierarchical multimodal ensemble learning framework specifically designed for social media popularity prediction. Our approach fundamentally differs from existing methods by implementing a three-tier hierarchical fusion architecture that progressively integrates features across multiple abstraction levels. The framework combines visual representations extracted from CLIP encoders, textual embeddings from transformer-based models [29], and comprehensive temporal-spatial metadata with user characteristics through a sophisticated hierarchical ensemble strategy. The key innovation of HyperFusion lies in its ability to automatically learn the optimal fusion weights at different hierarchical levels while simultaneously addressing the data scarcity problem through a two-stage training methodology incorporating pseudo-labeling and iterative refinement. Additionally, we introduce novel cross-modal similarity measures and hierarchical clustering features that effectively capture inter-modal dependencies, enabling more nuanced understanding of content popularity patterns.

Our comprehensive experimental evaluation on the SMP Challenge dataset [35] demonstrates that HyperFusion achieves competitive performance, ranking **third place** in the Image Track. This result provides valuable insights into the relative importance of different modalities and feature types in determining social media content popularity. The hierarchical fusion approach consistently outperforms traditional concatenation methods, validating the effectiveness of our proposed architecture in handling complex multimodal interactions.

2 RELATED WORK

Social media popularity prediction has evolved from early hand-crafted feature approaches to sophisticated multimodal learning [10, 15, 19, 27] frameworks. Initial studies [21, 37] focused on extracting features from textual descriptions, temporal patterns, and user demographics, establishing the fundamental understanding that engagement prediction requires comprehensive modeling of heterogeneous information sources. The transition to deep learning methodologies has fundamentally transformed this landscape, with contemporary approaches [8, 9, 20] leveraging powerful vision models [25, 26, 28, 39] or vision-language models such as CLIP and BERT [5] to extract rich semantic representations from both visual and textual content. These pre-trained encoders enable unified understanding of multimodal posts, capturing subtle engagement cues that traditional feature engineering methods often overlook.

The integration of multimodal information has become increasingly sophisticated through advanced fusion architectures. Hierarchical and attention-based mechanisms have shown particular effectiveness in modeling complex cross-modal interactions [30], while ensemble learning strategies have emerged as essential components of state-of-the-art solutions. Modern frameworks typically

combine gradient boosting methods including XGBoost [2], CatBoost [6], and LightGBM [13] with neural network architectures such as TabNet [1] and multilayer perceptrons, capitalizing on the complementary strengths of tree-based models in handling heterogeneous features and neural networks in capturing non-linear relationships. In addition to architectural improvements, data-centric strategies have also gained attention. To address data scarcity challenges inherent in social media datasets, recent works [12, 16, 18] have increasingly adopted pseudo-labeling strategies and iterative training procedures that effectively utilize unlabeled data while improving model robustness.

Despite these advances, several fundamental challenges persist in social media popularity prediction. The highly skewed distribution of engagement scores, noisy or incomplete metadata, and the dynamic nature of user preferences continue to complicate model development and evaluation. These challenges have driven the development of more robust feature extraction pipelines, sophisticated ensemble methodologies, and adaptive training strategies that can handle the inherent uncertainty and variability in social media data. The convergence of these methodological advances has established a clear paradigm emphasizing multimodal feature integration, powerful pre-trained representation models, and ensemble learning strategies as the foundation for effective social media popularity prediction systems.

3 METHODOLOGY

3.1 Overview

In this section, we present the design of the HyperFusion framework, which systematically addresses the challenge of predicting social media popularity through comprehensive multimodal analysis. The architecture of HyperFusion is specifically tailored to integrate heterogeneous sources of information and model complex interactions across multiple modalities. The pipeline begins with the extraction of high-level semantic representations from visual content using pretrained encoders, followed by the integration of textual features, user characteristics, and spatiotemporal signals. To ensure feature compatibility and minimize the influence of noise, we implement a series of preprocessing steps, including normalization, outlier removal, and dimensionality reduction. The resulting multimodal feature set is subsequently processed through an ensemble of specialized models, each capable of capturing distinct aspects of the prediction task. The overall framework is designed to be robust and interpretable, facilitating reliable modeling of the diverse and dynamic nature of social media content. In the following subsections, we describe each component of the pipeline in detail.

3.2 Problem Formulation

We formulate social media popularity prediction as a multimodal regression problem that leverages comprehensive feature representations. Given a social media post p with associated metadata, our objective is to predict its popularity score $y \in \mathbb{R}^+$ through the mapping function:

$$y = f(\mathbf{v}, \mathbf{t}, \mathbf{s}, \mathbf{u}, \mathbf{r}; \theta) \quad (1)$$

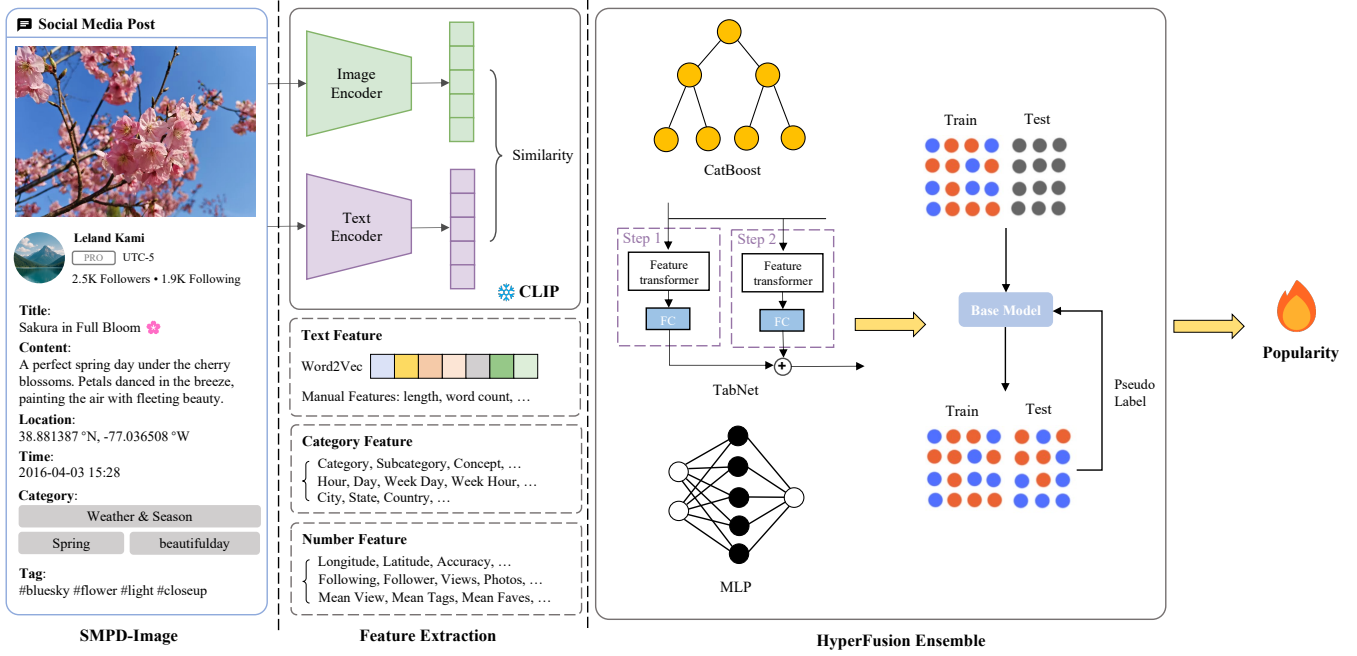


Figure 1: The HyperFusion framework pipeline integrates visual, textual, spatiotemporal, and user features through hierarchical ensemble learning for multimodal prediction of social media content popularity.

where $\mathbf{v} \in \mathbb{R}^{d_v}$, $\mathbf{t} \in \mathbb{R}^{d_t}$, $\mathbf{s} \in \mathbb{R}^{d_s}$, $\mathbf{u} \in \mathbb{R}^{d_u}$, and $\mathbf{r} \in \mathbb{R}^{d_r}$ represent visual, textual, spatiotemporal, user, and cross-modal similarity feature vectors respectively, with θ denoting the learnable parameters of the prediction model.

3.3 Multimodal Feature Construction and Integration

A central aspect of the HyperFusion framework is the systematic extraction and integration of multimodal features that capture the diverse factors influencing social media popularity. We organize feature engineering into five principal components: visual features, textual features, spatiotemporal features, user characteristics, and cross-modal coherence measures, each offering distinct perspectives on content engagement potential.

Visual Feature Extraction. We employ a pretrained CLIP encoder to extract deep visual representations from each image. The CLIP visual encoder processes input images to generate high-dimensional embeddings that capture semantic visual content. Specifically, let $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ denote an input image, where H and W represent height and width respectively. The visual feature \mathbf{v}_{visual} is computed as:

$$\mathbf{v}_{visual} = \text{CLIP}_{visual}(\mathbf{I}) \quad (2)$$

where $\text{CLIP}_{visual}(\cdot)$ represents the CLIP visual encoder that outputs a 512-dimensional feature vector. To further enhance representational capacity while maintaining computational efficiency, we apply Principal Component Analysis (PCA) [7] for dimensionality optimization when necessary.

Textual Feature Extraction. Our textual feature extraction combines multiple approaches to capture both semantic and statistical properties of text content. We utilize CLIP text encoders for cross-modal semantic alignment and GloVe [22] embeddings for traditional word-level representations. For hashtags and titles, we generate 300-dimensional feature vectors using GloVe embeddings. Let $\mathbf{T} = \{w_1, w_2, \dots, w_n\}$ represent a text sequence with n words. The textual feature $\mathbf{v}_{textual}$ is constructed as:

$$\mathbf{v}_{textual} = [\text{CLIP}_{text}(\mathbf{T}); \text{GloVe}(\mathbf{T}); \mathbf{f}_{stat}(\mathbf{T})] \quad (3)$$

where $\text{CLIP}_{text}(\cdot)$ generates semantic embeddings, $\text{GloVe}(\cdot)$ produces word-level representations, and $\mathbf{f}_{stat}(\cdot)$ extracts statistical features including text length, word count, and linguistic patterns.

Spatiotemporal and User Modeling. Temporal dynamics are captured through posting timestamps, account age calculations, and seasonal patterns that reflect content virality trends. Geographic context is encoded via coordinate information and location accuracy measures. User characteristics are modeled through behavioral patterns, engagement history, and professional status indicators. We apply Singular Value Decomposition (SVD) to user interaction matrices to generate compact user and location embeddings. Formally, given a user-item interaction matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$, the SVD decomposition yields:

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (4)$$

where $\mathbf{U} \in \mathbb{R}^{m \times k}$ provides 399-dimensional user embeddings and location embeddings of 400 dimensions are derived from $\mathbf{V} \in \mathbb{R}^{n \times k}$, capturing latent behavioral and geographic patterns.

Cross-Modal Coherence Measurement. To quantify the semantic alignment between visual and textual content, we compute

CLIP-based similarity scores that measure multimodal consistency. Given visual features \mathbf{v}_{visual} and textual features $\mathbf{v}_{textual}$, the cross-modal similarity is computed as:

$$s_{cross} = \frac{\mathbf{v}_{visual} \cdot \mathbf{v}_{textual}}{\|\mathbf{v}_{visual}\| \|\mathbf{v}_{textual}\|} \quad (5)$$

These coherence measures provide crucial insights into content consistency, which significantly influences user engagement and sharing behavior.

Feature Integration and Preprocessing. We concatenate all extracted features to form a unified multimodal feature vector for each post. Formally, let \mathbf{v}_{visual} , $\mathbf{v}_{textual}$, $\mathbf{v}_{spatial}$, \mathbf{v}_{user} , and \mathbf{v}_{cross} represent the feature vectors from different modalities. The final multimodal feature vector \mathbf{x} is constructed as:

$$\mathbf{x} = [\mathbf{v}_{visual}; \mathbf{v}_{textual}; \mathbf{v}_{spatial}; \mathbf{v}_{user}; \mathbf{v}_{cross}] \quad (6)$$

where $[\cdot; \cdot]$ denotes the concatenation operation. Before feeding the features into the models, we implement comprehensive preprocessing including missing value imputation with context-appropriate defaults, outlier filtering using the interquartile range method, and feature normalization to ensure stable distributions across different modalities.

3.4 HyperFusion Ensemble Architecture

Our HyperFusion framework implements a sophisticated ensemble strategy that combines multiple specialized models, each designed to capture different aspects of the multimodal feature space and prediction task.

Model Components. The ensemble incorporates four complementary architectures, each optimized for specific characteristics of the multimodal data. CatBoost Regressor serves as the primary gradient boosting component, specifically designed for structured data with categorical features. It utilizes category-aware encoding strategies and handles missing values naturally. TabNet provides interpretable deep tabular learning with sequential attention mechanisms that enable automatic feature selection and provide insights into feature importance. A custom Multi-Layer Perceptron (MLP) with co-attention layers captures complex visual-textual interactions through learned attention weights, enabling fine-grained cross-modal understanding. The CLIP-based Hierarchical Predictor leverages pretrained multimodal representations through multiple fully connected layers with ReLU activation and batch normalization.

Training Objective and Optimization. For the ensemble training, we employ a robust loss function that balances prediction accuracy with outlier resistance. The training objective for each model f_i is formulated using the Huber loss:

$$\mathcal{L}_s(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{if } |y - \hat{y}| \leq \delta \\ \delta|y - \hat{y}| - \frac{1}{2}\delta^2, & \text{otherwise} \end{cases} \quad (7)$$

where y denotes the ground-truth popularity score, \hat{y} represents the predicted value, and δ determines the transition point between quadratic and linear behavior, thereby improving robustness to outliers while maintaining sensitivity to regular samples.

Ensemble Integration Strategy. The final prediction combines individual model outputs through optimized weighted averaging. We apply five-fold cross-validation by splitting the training set

into five subsets and iteratively using four for training and one for validation. For each fold k , each model $f_i^{(k)}$ is trained independently. The ensemble prediction for fold k is computed as:

$$\hat{y}^{(k)} = \sum_{i=1}^N w_i^{(k)} \cdot f_i^{(k)}(\mathbf{x}) \quad (8)$$

The final prediction \hat{y} is calculated by averaging the outputs across all folds:

$$\hat{y} = \frac{1}{K} \sum_{k=1}^K \hat{y}^{(k)} \quad (9)$$

where $K = 5$ represents the number of folds, and weights $w_i^{(k)}$ are optimized through cross-validation to maximize ensemble performance on validation data.

3.5 Semi-Supervised Learning Strategy

We employ a sophisticated two-stage training approach that maximizes both model performance and generalization capability through iterative pseudo-labeling and semi-supervised learning techniques.

Base Model Training. In the initial stage, individual models are trained on the original labeled dataset using stratified cross-validation for robust hyperparameter optimization. Each model specializes in different aspects of the multimodal feature space: CatBoost focuses on categorical relationships and non-linear patterns inherent in user behavior and metadata, TabNet provides interpretable attention-based learning with automatic feature selection capabilities, and the MLP captures complex multimodal interactions through sophisticated co-attention mechanisms that model visual-textual dependencies.

Iterative Pseudo-Label Enhancement. The second stage implements an iterative pseudo-labeling strategy to effectively leverage unlabeled test data. High-confidence ensemble predictions are systematically incorporated as pseudo-labels for additional training iterations. This semi-supervised approach updates the training dataset by selecting predictions with confidence scores above empirically determined thresholds τ , defined as:

$$\tau = \mu_{confidence} + \alpha \cdot \sigma_{confidence} \quad (10)$$

where $\mu_{confidence}$ and $\sigma_{confidence}$ represent the mean and standard deviation of prediction confidence scores, and α is a hyperparameter controlling the selection strictness. This strategy enables the model to learn from broader data distributions while maintaining prediction quality and avoiding the incorporation of low-quality pseudo-labels that could degrade performance.

4 EXPERIMENT

4.1 Dataset

The SMP dataset comprises 486,000 image posts from 70,000 users over a 16-month period, encompassing a wide range of content categories [32]. Each post is associated with high-resolution images, textual descriptions, posting time, geographic information, and user profiles. The dataset supports multimodal popularity prediction by providing aligned visual, textual, spatiotemporal, and user-related features. Popularity scores are continuous values derived from real user engagement, enabling regression-based modeling and evaluation.

4.2 Evaluation Metrics

To comprehensively evaluate model performance, we adopt two metrics: Spearman’s Rank Correlation (SRC) and Mean Absolute Error (MAE). SRC assesses the monotonic relationship between predicted and actual popularity, reflecting the model’s ability to preserve correct ranking among samples. The SRC is defined as:

$$\text{SRC} = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (11)$$

where d_i is the difference between the ranks of the i -th predicted and ground-truth values. MAE measures the average absolute difference between predicted and ground-truth values, providing an interpretable error in the original scale:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (12)$$

This dual-metric evaluation ensures both ranking fidelity and numerical accuracy are considered in the assessment of popularity prediction models.

4.3 Main Results

4.3.1 Overall Performance. Our model achieved strong results on the official evaluation dataset, with a Spearman’s rank correlation coefficient (SRC) of 0.7324 and a mean absolute error (MAE) of 1.2402. These results reflect the effectiveness of our multimodal feature fusion and model design. The approach generalizes well across diverse samples and demonstrates robust performance in popularity prediction tasks.

4.3.2 Ablation Study. To further elucidate the contribution of each modality and the ensemble strategy, we conduct a focused ablation study. [Table 1](#) summarizes the results for the following configurations: the full model, and variants with each major modality or the ensemble mechanism removed. The evaluation is performed on the official validation set, and both SRC and MAE are reported for each setting.

Table 1: Ablation study results on the validation set.

Methods	SRC	MAE
Full Model	0.7324	1.2402
w/o Visual	0.6916	1.3503
w/o Metadata	0.7068	1.3441
w/o Textual	0.7196	1.2663
w/o Geolocation	0.7294	1.2445
w/o Ensemble	0.7253	1.2631
w/o Pseudo Label	0.7287	1.2774
w/o Filter	0.7249	1.2785
w/o 5-Fold	0.7290	1.2773

Our full model achieves the best performance, confirming the benefits of integrating multiple modalities and advanced training strategies. We observe a significant performance drop when removing visual features, which underscores the importance of visual content in predicting popularity. Similarly, excluding textual and metadata features also degrades performance, highlighting their

complementary roles in providing semantic and contextual information. The removal of geolocation data results in a minor accuracy decrease, suggesting its auxiliary yet valuable contribution.

We also investigate the impact of our proposed training techniques. Removing the pseudo-labeling stage leads to a noticeable performance decline, which validates the effectiveness of our semi-supervised approach in leveraging unlabeled data. The exclusion of the outlier filtering process also impairs results, confirming that robust data preprocessing is critical for model stability. Furthermore, forgoing the 5-fold cross-validation strategy for a single training split results in a substantial performance drop, which highlights the importance of cross-validation for mitigating overfitting and enhancing generalization. These results collectively affirm that each component of our framework, from feature engineering to training methodology, is integral to achieving strong predictive performance.

4.3.3 Feature Importance. To further elucidate the predictive mechanism of our model, we conduct a comprehensive feature importance analysis based on the averaged results across multiple runs. [Table 2](#) presents the top 20 most influential features. User-centric attributes, such as user identity, posting activity, and follower statistics, consistently dominate the ranking, highlighting the pivotal role of user information in popularity prediction. Semantic features, including concept category and tag-related statistics, also contribute substantially, reflecting the importance of content semantics. Visual and temporal cues, as well as latent representations from user and tag embeddings, provide complementary perspectives that enhance model robustness. The diversity among the top features underscores the necessity of integrating heterogeneous modalities within a unified framework.

Table 2: Importance of different features.

Rank	Feature	Importance	Rank	Feature	Importance
1	User ID	8.85	11	Total Tags	0.67
2	Concept	6.80	12	Group Count	0.63
3	Post Count	4.81	13	Mean Favorites	0.56
4	Follower Count	4.33	14	Tag Vec 218	0.52
5	Total Views	3.28	15	Tag Vec 15	0.51
6	Photo Count	2.66	16	User Vec 18	0.51
7	Tag Number	2.25	17	First Week Taken	0.48
8	First Date Posted	0.89	18	User Vec 16	0.43
9	Mean Tag	0.84	19	Following Count	0.42
10	First Date Taken	0.80	20	Tag Vec 224	0.42

These findings confirm that user-centric and semantic features are indispensable for accurate popularity estimation, while visual and temporal signals provide valuable complementary information. The integration of diverse modalities enables the model to capture the multifaceted nature of social media dynamics, leading to robust and generalizable predictions.

4.4 Analysis

To gain deeper insight into the predictive behavior of our model, we analyze the distribution of predicted and true popularity scores on the validation set. As illustrated in [Figure 2](#), both the histogram and density curves reveal that the predicted scores closely follow the empirical distribution of the ground-truth labels. The model

successfully captures the unimodal and right-skewed nature of the data, with the highest density concentrated between 5 and 10. This alignment demonstrates that the model is able to learn the dominant statistical patterns present in real-world popularity data.

A moderate smoothing effect is observed in the predicted distribution, which can be attributed to the ensemble inference and regularization strategies that promote generalization. While the model achieves reliable calibration across the majority of the label range, there is a tendency for predictions to be more conservative at the lower and upper extremes. In particular, the frequency of predicted scores in the lowest and highest intervals is slightly reduced compared to the true distribution, reflecting the challenge posed by data imbalance in these regions.

Despite these limitations, the overall consistency between predicted and actual distributions underscores the robustness of the proposed approach for large-scale popularity estimation. The model produces scores with meaningful variance and stable calibration, supporting its practical applicability in realistic social media scenarios. Future work may explore advanced techniques such as label distribution adjustment or targeted loss re-weighting to further improve performance on underrepresented cases.

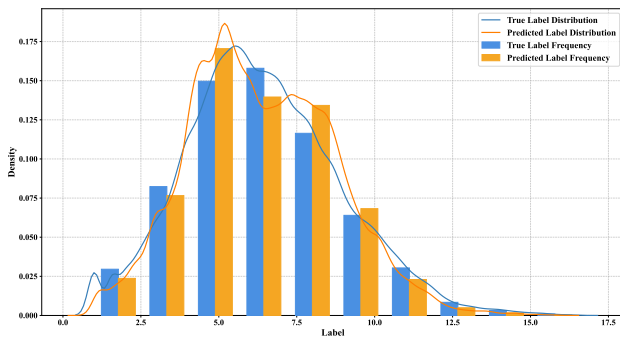


Figure 2: Histogram and kernel density estimation of predicted vs. ground-truth label distributions on the validation set.

5 CONCLUSION AND FUTURE WORK

This work presents HyperFusion, a comprehensive hierarchical multimodal ensemble learning framework for social media popularity prediction that systematically integrates visual, textual, spatiotemporal, and user behavioral information. Through careful orchestration of CLIP-based visual encoders, transformer-derived textual embeddings, and sophisticated ensemble strategies, our approach effectively captures the complex multimodal patterns underlying social engagement dynamics. The experimental evaluation demonstrates that the proposed hierarchical fusion architecture achieves competitive performance with an SRC of 0.7324 and MAE of 1.2402, validating the effectiveness of progressive feature integration across multiple abstraction levels.

The comprehensive ablation studies reveal that each modality contributes meaningfully to the overall predictive performance, with visual features providing essential aesthetic and semantic

cues, textual components capturing linguistic patterns and content semantics, and user characteristics offering crucial behavioral insights. The feature importance analysis further confirms that user-centric attributes dominate the predictive landscape, while semantic features and cross-modal coherence measures provide valuable complementary information. These findings highlight the importance of comprehensive multimodal modeling for accurate popularity prediction.

Nonetheless, several challenges remain for future exploration. The dynamic nature of social media trends, potential semantic inconsistencies across modalities, and the computational complexity of hierarchical fusion present ongoing research opportunities. Future investigations may focus on developing more adaptive fusion mechanisms that can dynamically adjust to evolving content patterns, as well as exploring temporal modeling approaches that capture the evolution of popularity over time. Moreover, enhancing the interpretability of cross-modal interactions and developing more efficient training strategies for large-scale deployment represent promising research directions.

In summary, this study contributes to the understanding of hierarchical multimodal learning in social media analysis and offers insights that may inform the development of more effective content popularity prediction systems across digital platforms.

References

- [1] Sercan Ö Arik and Tomas Pfister. 2021. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 6679–6687.
- [2] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.
- [3] Xiaolu Chen, <https://orcid.org/0000-0001-5819-5371>, View Profile, Weilong Chen, <https://orcid.org/0000-0003-2202-601X>, View Profile, Chenghao Huang, <https://orcid.org/0000-0003-2961-6282>, View Profile, Zhongjian Zhang, <https://orcid.org/0009-0000-6180-4342>, View Profile, Lixin Duan, <https://orcid.org/0000-0002-0723-4016>, View Profile, Yanru Zhang, <https://orcid.org/0000-0003-4182-2150>, and View Profile. 2023. Double-Fine-Tuning Multi-Objective Vision-and-Language Transformer for Social Media Popularity Prediction. *Proceedings of the 31st ACM International Conference on Multimedia* (Oct. 2023), 9462–9466. doi:10.1145/3581783.3612845
- [4] Zhangtao Cheng, Jienan Zhang, Xovee Xu, Goce Trajcevski, Ting Zhong, and Fan Zhou. 2024. Retrieval-Augmented Hypergraph for Multimodal Social Media Popularity Prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*. Association for Computing Machinery, New York, NY, USA, 445–455. doi:10.1145/3637528.3672041
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.
- [6] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363* (2018).
- [7] Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* 24, 6 (1933), 417.
- [8] Chih-Chung Hsu, Chia-Ming Lee, Yu-Fan Lin, Yi-Shiuan Chou, Chih-Yu Jian, and Chi-Han Tsai. 2024. Revisiting Vision-Language Features Adaptation and Inconsistency for Social Media Popularity Prediction. *arXiv:2407.00556* doi:10.48550/arXiv.2407.00556
- [9] Wenhao Hu, Weilong Chen, Weimin Yuan, Yan Wang, Shimin Cai, and Yanru Zhang. 2024. Dual-Stream Pre-Training Transformer to Enhance Multimodal Learning for Social Media Prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. Association for Computing Machinery, New York, NY, USA, 11450–11456. doi:10.1145/3664647.3688998
- [10] Yangliu Hu, Zikai Song, Na Feng, Yawei Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2025. Sf2t: Self-supervised fragment finetuning of video-llms for fine-grained understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 29108–29117.

- [11] Dahyun Jeong, Hyelim Son, Yunjin Choi, and Keunwoo Kim. 2024. Enhancing social media post popularity prediction with visual content. *Journal of the Korean Statistical Society* 53, 3 (2024), 844–882.
- [12] Peipei Kang, Zehang Lin, Shaohua Teng, Guipeng Zhang, Lingni Guo, and Wei Zhang. 2019. Catboost-based framework with additional user information for social media popularity prediction. In *Proceedings of the 27th ACM international conference on multimedia*. 2677–2681.
- [13] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).
- [14] Xin Lai, Yihong Zhang, and Wei Zhang. 2020. HyFea: Winning Solution to Social Media Popularity Prediction for Multimedia Grand Challenge 2020. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. Association for Computing Machinery, New York, NY, USA, 4565–4569. doi:10.1145/3394171.3416273
- [15] Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. 2024. Coupled mamba: Enhanced multi-modal fusion with coupled state space model. *arXiv preprint arXiv:2405.18014* (2024).
- [16] Hung-Hsiang Lin, Jiun-Da Lin, Jose Jaena Mari Ople, Jun-Cheng Chen, and Kai-Lung Hua. 2021. Social media popularity prediction based on multi-modal self-attention mechanisms. *IEEE Access* 10 (2021), 4448–4455.
- [17] Yu-Shi Lin and Anthony J.T. Lee. 2024. MMF: Winning Solution to Social Media Popularity Prediction Challenge 2024. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. Association for Computing Machinery, New York, NY, USA, 11445–11449. doi:10.1145/3664647.3688997
- [18] Yu-Shi Lin and Anthony JT Lee. 2024. MMF: Winning Solution to Social Media Popularity Prediction Challenge 2024. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11445–11449.
- [19] Run Luo, Zikai Song, Lintao Ma, Jinlin Wei, Wei Yang, and Min Yang. 2024. Diffusiontrack: Diffusion model for multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 3991–3999.
- [20] Shijian Mao, Wudong Xi, Lei Yu, Gaotian Lü, Xingxing Xing, Xingchen Zhou, and Wei Wan. 2023. Enhanced catboost with stacking features for social media prediction. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9430–9435.
- [21] Mayank Meghawat, Satyendra Yadav, Debanjan Mahata, Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. 2018. A multimodal approach to predict social media popularity. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, 190–195.
- [22] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
- [23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmlR, 8748–8763.
- [24] Social Media Statistic [n.d.]. Global Social Media Statistics. <https://datareportal.com/social-media-users>.
- [25] Zikai Song, Run Luo, Lintao Ma, Ying Tang, Yi-Ping Phoebe Chen, Junqing Yu, and Wei Yang. 2025. Temporal Coherent Object Flow for Multi-Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 6978–6986.
- [26] Zikai Song, Run Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2023. Compact transformer tracker with correlative masked modeling. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 2321–2329.
- [27] Zikai Song, Ying Tang, Run Luo, Lintao Ma, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2024. Autogenic language embedding for coherent point tracking. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2021–2030.
- [28] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2022. Transformer tracking with cyclic shifting window attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8791–8800.
- [29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [30] Jing Wang, Shuo Yang, Hui Zhao, and Yue Yang. 2023. Social media popularity prediction with multimodal hierarchical fusion model. *Computer Speech & Language* 80 (2023), 101490.
- [31] Kai Wang, Penghui Wang, Xin Chen, Qiushi Huang, Zhendong Mao, and Yongdong Zhang. 2020. A Feature Generalization Framework for Social Media Popularity Prediction. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. Association for Computing Machinery, New York, NY, USA, 4570–4574. doi:10.1145/3394171.3416294
- [32] Bo Wu, Wen-Huang Cheng, Peiye Liu, Bei Liu, Zhaoyang Zeng, and Jiebo Luo. 2019. Smp challenge: An overview of social media prediction challenge 2019. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2667–2671.
- [33] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, Qiushi Huang, Jintao Li, and Tao Mei. 2017. Sequential Prediction of Social Media Popularity with Deep Temporal Context Networks. *arXiv:1712.04443 [cs]* doi:10.48550/arXiv.1712.04443
- [34] Bo Wu, Peiye Liu, Wen-Huang Cheng, Bei Liu, Zhaoyang Zeng, Jia Wang, Qiushi Huang, and Jiebo Luo. 2023. SMP Challenge: An overview and analysis of social media prediction challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9651–9655.
- [35] Bo Wu, Peiye Liu, Qiushi Huang, Zhaoyang Zeng, Jia Wang, Bei Liu, Jiebo Luo, and Wen-Huang Cheng. 2024. SMP Challenge Summary: Social Media Prediction Challenge. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11442–11444.
- [36] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. 2016. Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.
- [37] Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha. 2018. User-guided hierarchical attention network for multi-modal social image popularity prediction. In *Proceedings of the 2018 world wide web conference*. 1277–1286.
- [38] Zhizhen Zhang, Ruihong Qiu, and Xiaohui Xie. 2024. Contrastive Learning for Implicit Social Factors in Social Media Popularity Prediction. <https://arxiv.org/abs/2410.09345v1>.
- [39] Hang Zhou, Jiale Cai, Yuteng Ye, Yonghui Feng, Chenxing Gao, Junqing Yu, Zikai Song, and Wei Yang. 2025. Video anomaly detection with motion and appearance guided patch diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 10761–10769.

References

- [1] Sercan Ö Arik and Tomas Pfister. 2021. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 6679–6687.
- [2] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.
- [3] Xiaolu Chen, <https://orcid.org/0000-0001-5819-5371>, View Profile, Weilong Chen, <https://orcid.org/0000-0003-2202-601X>, View Profile, Chenghao Huang, <https://orcid.org/0000-0003-2961-6282>, View Profile, Zhongjian Zhang, <https://orcid.org/0009-0000-6180-4342>, View Profile, Lixin Duan, <https://orcid.org/0000-0002-0723-4016>, View Profile, Yanru Zhang, <https://orcid.org/0000-0003-4182-2150>, and View Profile. 2023. Double-Fine-Tuning Multi-Objective Vision-and-Language Transformer for Social Media Popularity Prediction. *Proceedings of the 31st ACM International Conference on Multimedia* (Oct. 2023), 9462–9466. doi:10.1145/3581783.3612845
- [4] Zhanqiao Cheng, Jienan Zhang, Xovee Xu, Goce Trajcevski, Ting Zhong, and Fan Zhou. 2024. Retrieval-Augmented Hypergraph for Multimodal Social Media Popularity Prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*. Association for Computing Machinery, New York, NY, USA, 445–455. doi:10.1145/3637528.3672041
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.
- [6] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363* (2018).
- [7] Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* 24, 6 (1933), 417.
- [8] Chih-Chung Hsu, Chia-Ming Lee, Yu-Fan Lin, Yi-Shuan Chou, Chih-Yu Jian, and Chi-Han Tsai. 2024. Revisiting Vision-Language Features Adaptation and Inconsistency for Social Media Popularity Prediction. *arXiv:2407.00556* doi:10.48550/arXiv.2407.00556
- [9] Wenhao Hu, Weilong Chen, Weimin Yuan, Yan Wang, Shimin Cai, and Yanru Zhang. 2024. Dual-Stream Pre-Training Transformer to Enhance Multimodal Learning for Social Media Prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. Association for Computing Machinery, New York, NY, USA, 11450–11456. doi:10.1145/3664647.3688998
- [10] Yangliu Hu, Zikai Song, Na Feng, Yawei Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2025. Sf2t: Self-supervised fragment finetuning of video-llms for fine-grained understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 29108–29117.
- [11] Dahyun Jeong, Hyelim Son, Yunjin Choi, and Keunwoo Kim. 2024. Enhancing social media post popularity prediction with visual content. *Journal of the Korean Statistical Society* 53, 3 (2024), 844–882.
- [12] Peipei Kang, Zehang Lin, Shaohua Teng, Guipeng Zhang, Lingni Guo, and Wei Zhang. 2019. Catboost-based framework with additional user information for social media popularity prediction. In *Proceedings of the 27th ACM international conference on multimedia*. 2677–2681.
- [13] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting

- decision tree. *Advances in neural information processing systems* 30 (2017).
- [14] Xin Lai, Yihong Zhang, and Wei Zhang. 2020. HyFea: Winning Solution to Social Media Popularity Prediction for Multimedia Grand Challenge 2020. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. Association for Computing Machinery, New York, NY, USA, 4565–4569. doi:10.1145/3394171.3416273
 - [15] Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. 2024. Coupled mamba: Enhanced multi-modal fusion with coupled state space model. *arXiv preprint arXiv:2405.18014* (2024).
 - [16] Hung-Hsiang Lin, Jiun-Da Lin, Jose Jaena Mari Ople, Jun-Cheng Chen, and Kai-Lung Hua. 2021. Social media popularity prediction based on multi-modal self-attention mechanisms. *IEEE Access* 10 (2021), 4448–4455.
 - [17] Yu-Shi Lin and Anthony J.T. Lee. 2024. MMF: Winning Solution to Social Media Popularity Prediction Challenge 2024. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. Association for Computing Machinery, New York, NY, USA, 11445–11449. doi:10.1145/3664647.3688997
 - [18] Yu-Shi Lin and Anthony JT Lee. 2024. MMF: Winning Solution to Social Media Popularity Prediction Challenge 2024. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11445–11449.
 - [19] Run Luo, Zikai Song, Lintao Ma, Jinlin Wei, Wei Yang, and Min Yang. 2024. Diffusiontrack: Diffusion model for multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 3991–3999.
 - [20] Shijian Mao, Wudong Xi, Lei Yu, Gaotian Lü, Xingxing Xing, Xingchen Zhou, and Wei Wan. 2023. Enhanced catboost with stacking features for social media prediction. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9430–9435.
 - [21] Mayank Meghawat, Satyendra Yadav, Debanjan Mahata, Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. 2018. A multimodal approach to predict social media popularity. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, 190–195.
 - [22] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.
 - [23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmlR, 8748–8763.
 - [24] Social Media Statistic [n.d.]. Global Social Media Statistics. <https://datareportal.com/social-media-users>.
 - [25] Zikai Song, Run Luo, Lintao Ma, Ying Tang, Yi-Ping Phoebe Chen, Junqing Yu, and Wei Yang. 2025. Temporal Coherent Object Flow for Multi-Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 6978–6986.
 - [26] Zikai Song, Run Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2023. Compact transformer tracker with correlative masked modeling. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 2321–2329.
 - [27] Zikai Song, Ying Tang, Run Luo, Lintao Ma, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2024. Autogenic language embedding for coherent point tracking. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2021–2030.
 - [28] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2022. Transformer tracking with cyclic shifting window attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8791–8800.
 - [29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
 - [30] Jing Wang, Shuo Yang, Hui Zhao, and Yue Yang. 2023. Social media popularity prediction with multimodal hierarchical fusion model. *Computer Speech & Language* 80 (2023), 101490.
 - [31] Kai Wang, Penghui Wang, Xin Chen, Qiushi Huang, Zhendong Mao, and Yongdong Zhang. 2020. A Feature Generalization Framework for Social Media Popularity Prediction. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. Association for Computing Machinery, New York, NY, USA, 4570–4574. doi:10.1145/3394171.3416294
 - [32] Bo Wu, Wen-Huang Cheng, Peiye Liu, Bei Liu, Zhaoyang Zeng, and Jiebo Luo. 2019. Smp challenge: An overview of social media prediction challenge 2019. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2667–2671.
 - [33] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, Qiushi Huang, Jintao Li, and Tao Mei. 2017. Sequential Prediction of Social Media Popularity with Deep Temporal Context Networks. *arXiv:1712.04443 [cs]* doi:10.48550/arXiv.1712.04443
 - [34] Bo Wu, Peiye Liu, Wen-Huang Cheng, Bei Liu, Zhaoyang Zeng, Jia Wang, Qiushi Huang, and Jiebo Luo. 2023. SMP Challenge: An overview and analysis of social media prediction challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9651–9655.
 - [35] Bo Wu, Peiye Liu, Qiushi Huang, Zhaoyang Zeng, Jia Wang, Bei Liu, Jiebo Luo, and Wen-Huang Cheng. 2024. SMP Challenge Summary: Social Media Prediction Challenge. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 11442–11444.
 - [36] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. 2016. Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.
 - [37] Wei Zhang, Wen Wang, Jun Wang, and Hongyuan Zha. 2018. User-guided hierarchical attention network for multi-modal social image popularity prediction. In *Proceedings of the 2018 world wide web conference*. 1277–1286.
 - [38] Zhizhen Zhang, Ruihong Qiu, and Xiaohui Xie. 2024. Contrastive Learning for Implicit Social Factors in Social Media Popularity Prediction. <https://arxiv.org/abs/2410.09345v1>.
 - [39] Hang Zhou, Jiale Cai, Yuteng Ye, Yonghui Feng, Chenxing Gao, Junqing Yu, Zikai Song, and Wei Yang. 2025. Video anomaly detection with motion and appearance guided patch diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 10761–10769.